Webからの関連語抽出の役割分類を用いた拡張と精度向上

白石　卓也† Bollegala Danushka† 石塚　満‡
東京大学工学部電子情報工学科†東京大学院情報理工学系研究科‡ 東京大学院情報理工学系研究科
電子情報学専攻　　電子情報学専攻‡

1. はじめに

シソーラスやオントロジーはWebに限らず，広く自然言語処理一般での重要な知識ベースである．新聞記事などをコーパスとする関連語(同義語，上位語，下位語，全体語，部分語)抽出の手法が今までに数多く提案されてきた．また，近年のWebの爆発的な普及によって，Web上の文章をコーパスとして用いる手法も提案されている．Webを利用する利点は，新語や意味の変化を追うことができる点にある．Webの情報は膨大であり，効率よく情報を収集するためにはWeb検索エンジンを利用することは必須である．

本論文では，Web検索エンジンを用いた既存の関連語抽出手法を応用し，特に同義語抽出の精度向上のための手法を提案する．

2. 関連研究

Webの発展によって，Web上の文章をコーパスとして関連語を抽出する様々な手法が提案されている．

[Bollegala,07]では，ワードペア間の類似度の測定を，Web検索のヒット件数やスニペットを用いて行う手法が提案され，その成果を受けた[渡部,09]はBollegalaの手法を関連語抽出に適用した．

渡部の手法は，まず，求めたい関係(同義関係や上位/下位関係)を満たすワードペアを多数用意して，それを用いたクエリでWeb検索を行い，得られたスニペットから関連語抽出に有意なレキシカルパターンを生成するというものである．

生成されたレキシカルパターンからクエリを作成してWeb検索を行い，スニペットから関連語候補を抽出する．例えば，request"dog"の同義語を得たい場合，同義関係を満たすワードペア群によって生成されたパターンである"X synonyms Y"などのXに"dog"を代入し，"dog synonyms *"というクエリを作成してWeb検索を行う．得られるスニペットに出現する表層的な単語の並びから，"*"に相当する部分を同義語候補として抽出できる．

渡部の継続研究として千葉は，requestの語から求められた同義語候補の中で高スコアの1語をピックアップしてanchorとし，anchorについても同義語抽出を行い，2語に共通する同義語候補のスコアを高くすることによってrequestの同義語抽出の精度向上を狙った．この手法は，2語に共通する語義に関してはスコアが高くなるが，共通しない同義語のスコアが低くなってしまった．

本論文では，千葉の手法の欠点を補った上で，さらに〈役割〉による分類を行い，同義語抽出の精度向上手法を提案する．

3. 手法

3. 1 オーバーラップによる分類

requestには多義性があり，様々な語義で同義語が異なっている．異なる2語の同義語候補のオーバーラップ(以下OLと表記する)は，1語だけの同義語抽出において多く含まれていたノイズが減少するという利点はあるが，多義性が無視されてしまう，という欠点がある．

そこで，OLに使用する語を複数用意する．それぞれが異なる語義でのrequestの同義語であれば，OLはそれぞれの語義での，ノイズの減少したrequestの同義語候補となる．複数のOLを再集合すれば，多義性を保ったままノイズの減少した同義語候補が得られると考えられる．

OLに使用する語をanchorと呼ぶことにし，それらを集めてanchorリストを作成する．anchorリストには，多様な語義でのrequestの同義語が精度良く含まれることが求められる．

以下の操作を繰り返す．
1.anchorリストからスコアの高いanchorを選ぶ．
2.anchorの同義語候補を求め，requestの同義語候補とのOLを作る．
3.OLにrequestが含まれるのなら，正しいOLとして採用する．
4.anchorリストからOLに含まれる語を削除．
5.anchorリストが空なら終了．

OLにはrequestとanchorに共通する語義での同義語が精度良く含まれているため，複数のOLを作成したことで「同義語候補の語義に関する大まかな分類ができた」とみなせる．

3. 2 〈役割〉による分類

正解データとして用いるRoget's II[1]では，"horse"には"equine species"や"hoofed mammal"といったエントリがあり，それらは異なるsynonymをもっていることになっている．これは「馬」という語義に関するエントリが2つ以上に分かれているということである．この1つのように見える語義から分割されたエントリを本論文では〈役割〉と呼ぶことにする．

〈役割〉は[Hearst,92]などの上位語抽出のパターンを用いて，Web検索のスニペットから具体的な名詞として抽出できる．ここでは，OLの同義語候補をさらに洗練するため，requestとanchorが共通してもっている〈役割〉によって分類する．

以下のような操作を各OLについて行う
1.anchorの〈役割〉抽出
"<anchor> or other *"というクエリでWeb検索を行い，スニペットから"*"に相当する部分を〈役割〉として抽出する．[2]

Sophisticate the related terms using Web search engine
†Takuya Shiraishi, Faculty of Engineering, University of Tokyo
†Bollegala DANUSHKA
†Mitsuru ISHIZUKA Graduate School of Information Science and Technology, University of Tokyo

[1]Roget'sII(http://thesaurus.reference.com/)
[2]YahooBoss(http://developer.yahoo.com/search/boss/)を用いる

2.〈役割〉の選択

抽出した〈役割〉のうち，requestにも当てはまるものを，上位/下位関係を満たすパターンに代入したクエリの検索結果によって選ぶ．同義語抽出は希少な語義や希少な語について扱うことが多く，あまりに形式的なパターンだとヒット件数が 0 になりやすい．そのため，今回は最も単純な"<request> is (a|an) <role>"というパターンを用いて Web 検索を行い，ヒット件数が H 件以上なら〈役割〉とみなす．また，希少な<役割>ほど重要であると仮定して，"* is (a|an) <role>"というクエリの検索ヒット件数の少ない方から N 語を分類に使う〈役割〉とする．

3.〈役割〉による分類

"<word> is (a|an) <role>"というクエリを OL に含まれる全ての語と全ての〈役割〉で作り Web 検索のヒット件数が H 件以上ならその〈役割〉のクラスに同義語候補を登録していく．

3．3 同義語候補のスコアリング

各同義語候補のスコアの基準は，以下の仮説にしたがっている．

1)希少な〈役割〉ほど，具体的な<役割>である．

2)具体的な〈役割〉を多く共有している語ほど，同義性が高い．

1)と 2)の仮説から，以下の式を同義語 c のスコアとした．

$$rScore(role)=\frac{C}{C+Hit(\text{* is a(an) <role>})}$$

$$cScore(c)=\sum_{role} rScore(role)\times\chi_{role}(c)$$

ただし，$\chi_{role}(c)=\begin{cases}1. \ if \quad c\in role\\ 0. \ if \quad c\notin role\end{cases}$

4．実験

表 1 に挙げられている 6 単語について，各手順での同義語の網羅性と最終的な精度を測定した．正解データは各語の Roget's II に名詞として登録されているエントリの synonym とした．

表 1 は各語の同義語候補の各段階での正解候補数と正解数と F 値の推移である．OL 前のデータは上位 100 個の抽出パターンによって抽出されたもので，anchor リストは OL 前のデータと同じものの中から上位 150 語を選んで作成し，〈役割〉分類でのパラメータは $H=1, N=10$ とした．

次に，スコアリングの定数を $C=100000$ で固定し，OL 前のデータ(bOL)，〈役割〉分類まで終わったがスコアは OL 前のデータと同じというデータ(nonScore)，スコアリング済みのデータ(Score)の 3 種類のデータについて，上位から同義語候補を取得していったときの適合率の推移を比較した．図 1

5．考察

Score の適合率の推移から分かるのは，〈役割〉の中にはノイズが一定の割合で混入してしまっているということである．is-a 関係を求めるために今回は"<word> is (a|an) <role>"というパターンを利用したが，is-a 関係には，厳密には"*(a|an)* <word> is (a|an) <role> "というパターンが必要である．しかしこのような形式的な表現は Web 検索エンジンではなかなかヒットしないという欠点がある．そのため，例えば"person is a magician"というようなクエリでも Web 検索がヒットしてしまい，"person"が"magician"に登録されてしまった．

また，nonScore と bOL にあまり差がないことは，〈役割〉による分類が，もともとのスコアが高い語については機能していないことを表している．

表 1．同義語候補の洗練と網羅性の推移

(size,ex)	OL 前	OL 後	最終	網羅性
magician	(781,49)	(316,46)	(269,43)	0.88
shore	(918,45)	(356,39)	(269,38)	0.84
slave	(973,31)	(364,28)	(296,24)	0.77
glass	(1000,31)	(373,28)	(309,28)	0.9
fruit	(815,22)	(310,20)	(273,20)	0.91
cord	(888,39)	(356,37)	(299,36)	0.92

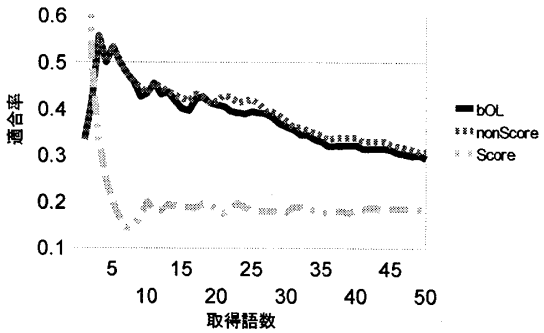

図 1．適合率の推移

これは，同義語と同族語とがうまく分離されていないことを意味している．本手法におけるこの問題の解決策として，今のところ 2 つの方法が考えられる．

1 つは，request と anchor に共通する〈役割〉だけでなく，共通しない〈役割〉も分類に使うということである．それをネガティブなクラスと考えて、マイナスのスコアとして計上すれば，同俗語はある程度スコアを低くでき，先の"person"のような一般的な語が"magician"に登録されてしまっても，一般的な語は他のネガティブな〈役割〉にも登録されるために結果としてそのスコアを下げることができるだろう．

もう 1 つは，同義語抽出パターンの性質をより詳しく明らかにすることで，現在の順位付けとは異なる指標を作ることである．例えば"X or Y"のような等位接続を含むパターンは同族性が高い語を抽出しやすいだろうとか，"X it means Y"のようなパターンは X の普段使われないような意味を説明しているので，どちらかといえば同義性の高い語を抽出しやすいのではないか，という具合に，レキシカルパターンは，適合率や再現率，F 値といった抽出精度に関する情報以外にも様々な情報を含んでいるはずである．これを定量的に表すことができれば，正解に加えるか加えないかがあいまいな同族語を，うまく扱えると考えられる．

6．まとめ

同義語抽出は希少な語や希少な語義を扱うことの多い，難しいタスクである．本論分ではオーバーラップを用いたり，〈役割〉ごとに同義語候補を分類していくことで同義語抽出の精度向上を試みた．

今後は同族語と同義語の区別を中心に精度を高めていく必要がある．

参考文献

[Hearst,92]M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539-545, 1992.

[Bollegala,07]D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th International World Wide Web Conference (WWW-07)*, pages 757-766, 2007.

[渡部,09]検索エンジンを用いた関連語の自動抽出 JSAI2009